Performance Analysis and Optimal Cooperative Cluster Size for Randomly Distributed Small Cells under Cloud RAN

Lei Zhang, Atta ul Quddus, Efstathios Katranaras, Dirk Wübben, Yinan Qi, Rahim Tafazolli

***Abstract*—One major advantage of cloud/centralized radio access network (C-RAN) is the ease of implementation of multi-cell coordination mechanisms to improve the system spectrum efficiency (SE). Theoretically, large number of cooperative cells lead to a higher SE, however, it may also cause significant delay due to extra channel state information (CSI) feedback and joint processing computational needs at the cloud data center, which is likely to result in performance degradation. In order to investigate the delay impact on the throughput gains, we divide the network into multiple clusters of cooperative small cells and formulate a throughput optimization problem. We model various delay factors and the sum-rate of the network as a function of cluster size, treating it as the main optimization variable. For our analysis, we consider both base stations' as well as users' geometric locations as random variables for both linear and planar network deployments. The output SINR (signal-to-interference-plus-noise ratio) and ergodic sum-rate are derived based on the homogenous Poisson point processing (PPP) model. The sum-rate optimization problem in terms of the cluster size is formulated and solved. Simulation results show that the proposed analytical framework can be utilized to accurately evaluate the performance of practical cloud-based small cell networks employing clustered cooperation.**

***Index Terms*—Cloud-RAN, CSI delay, latency, optimal cooperative cluster, Poisson point processing**

I. Introduction

As a promising candidate technology for next generation wireless communications, cloud (or centralized) radio access network (C-RAN) has drawn significant attention by both academia and industry in the last few years. Apart from C-RAN's advantage of reducing radio site operations and capital costs, another benefit is relate to ease in the implementation of multi-cell coordination mechanisms such as coordinated multi-point transmission and reception (CoMP) [1], [2], [3], [4], thus promising higher system performance through efficient interference management. In addition, the cloud-based architecture provides the flexibility of splitting the radio access functionalities between the cloud and the remote sites depending on the backhaul link capacity and software/hardware processing capability of the access and cloud entities in the network [2], [5], [6]. One of the most popular functional split options in a C-RAN is to consider a high computational capability central processor taking high-complexity tasks in the cloud, and a set of densely deployed, low power, low-complexity radio remote heads (RRHs) [7],[8]. This option can harness the benefit of deploying a low-cost dense small cell network, while at the same time efficient interference avoidance and cancelation algorithms across multiple small cells can be realized through centralized processing in order to improve network spectral efficiency (SE).

Theoretically, larger cooperation cluster size (i.e. number of cooperating cells) leads to better interference cancelation and higher system SE. However, this is in practice not true if real-world implementation factors, such as latency, are taken into account. Larger number of cooperating cells/antennas results in more complex channel estimation and precoding implementation; this is especially true for advanced channel estimators such as minimum mean square error (MMSE) [9], [10] and zero-forcing (ZF) precoders [11] whose complexity is in cubic-order of the number of involved (transmitting or receiving) antennas. In addition, more antennas/cells/users in a cluster imply more CSI required for precoding, bringing further CSI feedback delay into the system. Furthermore, due to general-purpose hardware processing in the cloud data center, and also due to the uncertainties in availability of computational resource, significant processing delay may get added. All these delays can cause mismatch between actual channels and the channel used for calculating precoder matrix, consequently, performance degradation results. Therefore, we conjecture that there must be an optimal cluster size, large enough to mitigate interference into a reasonable level yet small enough to save the performance loss due to the delay-caused channel mismatch.

To optimize the cluster size of cloud-based small cell networks, building a mathematic link between the optimizing criterion in terms of sum-rate and the cluster size is the key. Two problems arise in that building process: 1) how to model the signal model of the network as a function of the cluster size and 2) how to map the cluster size as a function of latency.

For the first problem, we start our analysis from the deployment of cellular system, where typically the base stations (BS) are fixed in homogeneous grid. However, the most significant change that has to be taken into consideration for cooperation in small cells (in comparison with a point-to-point MIMO system) is the geometric location randomness which leads to the uncertainty of the large-scale fading. Towards this end, several papers have considered large scale fading as well as small scale fading in analyzing the ergodic capacity of cooperative systems for uplink [12] and downlink [13], [14], [15], [16], considering BSs locations fixed while treating UEs locations as random variables. However, this model is likely to be inaccurate for heterogeneous networks consisting of small cell deployments both in urban and suburban areas, where cell radius varies significantly and should be modeled as a

random variable in itself. In a befitting direction, in [17], [18], an analysis was presented by introducing an extra source of randomness, i.e. modeling the position of the base station as a homogeneous Poisson point processing (PPP), which will be used as a framework in our analysis in Section III. In addition, a tractable model for non-coherent joint transmission base station cooperation is established and closed-form for SINR distribution by considering a single UE is proposed in [19].

Regarding the second problem of mapping cluster size with latency, few works in literature so far have considered the impact of latency in multi-cell cooperation systems. The authors in [20] and [21] considered CSI feedback delay in their analysis for distributed antenna systems. Moreover, backhaul latency models for various backhaul topologies and technologies were only introduced in [22] and [23], however, the performance analysis and proposed algorithms are based on a single cooperative cluster instead of a network composed of multiple clusters that may interfere to each other.

In this paper, we consider a set of clusters in a cloud-based network sharing the same cloud resources, where each cluster is composed by a number of RRHs performing joint processing and operating as multi-antenna BS. In our work, we consider both RRHs' and users' geometric location as random variables based on homogeneous PPP [17], [18] as well as the effects of processing and CSI feedback delay. This approach is not only more generic but also more realistic considering the dynamic deployment nature of small cells in the future. The output SINR is derived in terms of the RRH and UE density in the presence of delay-caused channel mismatch. All of the parameters are converted into a function of cluster size to formulate the optimization problem. We briefly summarize the contributions and constraints of our work on the cluster size optimization for cloud-based small cell networks as follows:

- By treating both UEs and RRHs locations as random variables, the path-loss in addition with the fast fading are considered in the channel model. In the presence of delay and given a specific cluster size, we derive the general output SINR expression in terms of node density ρ_{U} and ρ_{R} by considering two representative linear precoders: MRT (maximum ratio transmission) and ZF, respectively. The analysis is divided into two steps. The first one assumes UE is located at an arbitrary distance to the cluster center; and in the second step, the analysis is generalized by treating the arbitrary distance as a random variable. Both linear and planar deployment of cells is considered.
- Considering an FDD (frequency division duplex) system, we build a generic delay model for the cloud-based small cell networks comprising the computational processing delay at the cloud data center, CSI feedback delay due to the transmission capacity limit link from UE to RRH, channel estimation delay, the propagation delay and the backhaul latency due to the data exchanging between cloud and small cells. However, the model can be extended to TDD (time division duplex) system straightforwardly. We then model the precoder delay as a function of the cooperative cluster size to show its impact on the performance.
- We show that the ergodic output SINR can be expressed as a function of the cluster size. The geographic area of a network is considered to be divided into separate clusters and an optimization problem is formulated by expressing ergodic sum-rate in terms of the cooperative cluster size. Due to the complex relationship between the cost function and the cluster size, numerical methods are used to show its consistency to the simulation results. Both large-scale and small-scale factors are considered in the channel model, however, the shadowing factor of the large scale fading, is not considered for tractability of analysis.
- The paper focuses on one of the most popular functional split options in cloud-based architecture [2], consisting of low-complexity low-cost RRHs and a cloud data center that can take as much functionality implementation as possible with perfect (in terms of capacity) backhaul[1] between them. However, the general method and optimization could be adopted for any other functional split options depending on the backhaul capacity and other hardware constraints.

Notations: Vectors and matrices are denoted by uppercase and lowercase bold letters, where $\{\cdot\}^H, \{\cdot\}^T, \{\cdot\}^*$ stand for the Hermitian conjugate, transpose and conjugate operation, respectively. $\mathcal{E}\{\cdot\}$ denotes the expectation operation. We use $(\dot{O}, R)$ to denote a circle with radius R and its center at $\dot{O}$. $[\mathbf{A}]_k$ and $\|\mathbf{A}\|$ refer to the k-th diagonal element and the Frobenius norm of matrix $\mathbf{A}$, respectively.

II. System Model

As shown in Fig. 1, we consider the downlink of a network comprising a large number of RRHs and we suppose that each RRH is connected with the cloud data center through fiber or other capacity unlimited backhaul links, while there is no direct physical link between RRHs. Therefore each RRH in a cooperative cluster can only acquire a local CSI and global CSI is accumulated at the cloud by RRH feedback via the backhaul links. The precoding matrix calculations will be done in the cloud. However, there are two main options for the precoding implementation: a) the implementing at the cloud and then forwarding the precoded I/Q signals to individual RRH for transmission; b) cloud-assisted implementation at each individual RRH, i.e. the modulated I/Q signal (before precoding) and relevant precoding coefficients will be sent from the cloud to the each RRHs and the rest of physical layer processing will take place in RRH. There are pros and cons for each architecture [24]. In this paper, we focus our investigation on case a) only, which is a more popular cloud architecture.

To mitigate the expected delay under joint transmission operation (mainly due to CSI feedback and precoding matrix calculation), we need to divide the network into a set of clusters with each one consisting of reasonable number of

[1]Strictly speaking, the link between RRH and cloud is called fronthaul while the link between the cloud to the core network is called backhaul. However, in some studies [2], [3], the link between the UE and RRH is called backhaul in order to separate it from the access link.

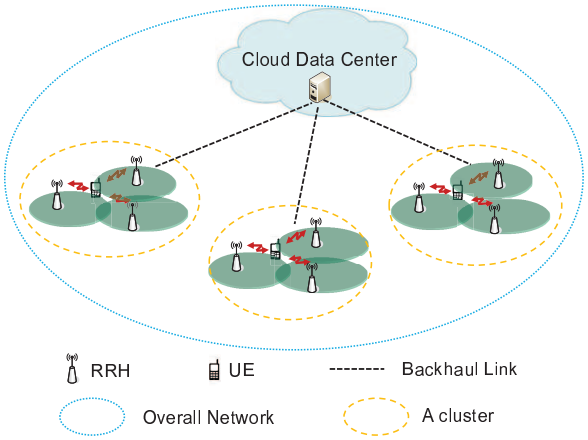

Fig. 1. Considered C-RAN architecture and RRH clustering.

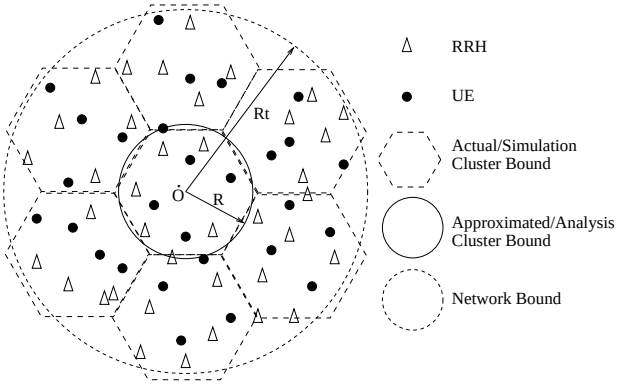

Fig. 2. Cluster division for planar deployment with each cluster being shaped as a hexagon.

cooperative RRHs as shown in Fig. 1. In this case, each cluster only requires CSI between the RRHs and the UEs within the cluster and the amount of feedback will drop since low dimension precoding requires less CSI. Meanwhile, precoding can be performed in the cloud separately for each cluster, therefore, the computational complexity will be reduced. Consequently, the introduced delays could be mitigated and performance loss due to the channel mismatch will be reduced. On the other hand, extra inter-cluster interference due to small cluster size may diminish throughput in each cluster. Therefore, there must be an optimal cluster size to trade off delay and interference for maximizing system performance.

A. Clustering Model

Consider a network served by a cloud in a d-dimensional space with a volume of V, where d could take the value of 1 or 2, corresponding to the linear or planar deployments, respectively. Here we assume the space is centrally symmetric and therefore the volume of the space could be generally expressed by $c_d R_t^d$, where c_d is the volume of the d-dimensional unit ball and R_t is the distance from the center to an arbitrary point on the bound of the d-dimensional space. Obviously, for linear and planar deployments, $c_1 = 2$ and $c_2 = \pi$, respectively.

Focusing on the planar deployment for demonstration purposes, we consider the network is divided into N_c *same area and same shape* clusters, i.e. each cluster has an area of $v = V/N_c$. One practical cluster shape in order to avoid adjacent cluster overlap and to maximize the density of packing in this 2D space is the hexagonal shape, as shown in Fig. 2. However, since hexagonal boundary is relatively difficult to analyze, the hexagonal cluster can be replaced by an equivalent circular cluster having the same central point and the same area as the hexagon. The approximation is practical and accurate since the sum-rate contribution from the edge of a cluster is marginal[2] (see Fig. 2). In addition, we consider a network comprising of b tiers clusters, i.e. the number of clusters in a network can only be $1 + 3b(b+1)$ [26], where b denotes the number of tiers. Fig. 2 gives an example of 1-tier network comprising 7 clusters. Note that under the assumption of network bounding with a circular one, the network is not completely overlapped by the clusters at the edge of the network. However, larger number of tiers (i.e. more clusters) leads to a negligible model error. In addition, the approximation is made at cluster edges, which contributes marginal sum-rate to the network[3]. Note that the above approximation for the planar deployment is not required for the linear case, where each cluster evenly occupies the same portion (length) of the network.

Next, we consider the active RRHs and UEs randomly scattered into the d-dimensional space with density ρ_{U} and ρ_{R}, respectively (see Fig. 2). We assume that each RRH is equipped with M antenna and each UE is equipped with single antenna. Thus, the number of UEs $n_{\mathrm{U},i}$, RRHs $n_{\mathrm{R},i}$ and transmission antennas $n_{\mathrm{A},i} = M n_{\mathrm{R},i}$ in each cluster (for $i = 1, 2, \cdots, N_c$) are variables. Since the UEs and RRHs are dropped based on the homogeneous PPP model, we can express the expected number of RRHs N_{R}, UEs N_{U}, and total transmission antennas N_{A} in any cluster i with volume v as [18]:

$$\begin{aligned} N_{\mathrm{U}} &= \mathcal{E}(n_{\mathrm{U},i}) = v\rho_{\mathrm{U}} \ , \quad N_{\mathrm{R}} = \mathcal{E}(n_{\mathrm{R},i}) = v\rho_{\mathrm{R}} \ , \\ N_{\mathrm{A}} &= \mathcal{E}(n_{\mathrm{A},i}) = \mathcal{E}(M n_{\mathrm{R},i}) = Mv\rho_{\mathrm{R}} \ , \end{aligned} \tag{1}$$

Apparently, the expected number of UEs and RRHs in the region outside of the considered cluster i is $(V - v)\rho_{\mathrm{U}}$ and $(V - v)\rho_{\mathrm{R}}$, respectively.

B. Channel Model

By considering both path-loss and small scale fast fading effects of the wireless access channel between RRH and UE, the baseband channel at time index t can be expressed as

$$\bar{\mathbf{h}}_{i,j,k}[t] = \mathbf{h}_{i,j,k}[t]\mathbf{G}_{i,j,k} \ , \tag{2}$$

where $\bar{\mathbf{h}}_{i,j,k}[t] \in \mathbb{C}^{1\times n_{\mathrm{A},j}}$ is the compound channel between the j-th RRH and k-th UE in the i-th cluster. $\mathbf{G}_{i,j,k} = \mathrm{diag}[g_{i,j,k,1}, g_{i,j,k,2}, \cdots, g_{i,j,k,n_{\mathrm{A},j}}] \in \mathbb{R}^{n_{\mathrm{A},j}\times n_{\mathrm{A},j}}$ is a diagonal

[2]Indeed, we adopt hexagonal cluster as the practical deployment in our simulations, the results will show the approximation error is negligible comparing with the analyzed results that are based on a circle shaped cluster.

[3]Our simulation results (as illustrated in Section V-1) show that this approximation provides valid results even with small number of tiers.

matrix with $g^2_{i,j,k,l}$ corresponding to the channel path-loss. $\mathbf{h}_{i,j,k}[t] \in \mathbb{C}^{1\times n_{\mathrm{A},j}}$ is the corresponding fast fading factor of the channel with complex Gaussian distribution $\mathcal{CN}(0,1)$.

The precoding matrix calculation is based on a delayed (outdated) version of the channel, at time $t-\triangle t$, which can be written as:

$$\bar{\mathbf{h}}_{i,j,k}[t-\triangle t] = \mathbf{h}_{i,j,k}[t-\triangle t]\mathbf{G}_{i,j,k}\ , \tag{3}$$

where we have assumed that RRHs and UEs are essentially static and the path-loss $\mathbf{G}_{i,j,k}$ is kept as a constant during the $\triangle t$ period.

Here we suppose that the channel is spatially uncorrelated but time correlated as follows [25]:

$$\mathbf{h}_{i,j,k}[t] = \mathbf{h}_{i,j,k}[t-\triangle t]\mathbf{\Lambda}_{i,j,k} + \mathbf{e}_{i,j,k}[t]\ , \tag{4}$$

where $\mathbf{\Lambda}_{i,j,k}$ is a diagonal matrix defined as $\mathbf{\Lambda}_{i,j,k} = \mathrm{diag}[\lambda_{i,j,k,1}, \lambda_{i,j,k,2}, \cdots, \lambda_{i,j,k,n_{\mathrm{A},j}}] \in \mathbb{R}^{n_{\mathrm{A},j}\times n_{\mathrm{A},j}}$ with $\lambda_{i,j,k,l} = \mathcal{J}_0(2\pi f_{D,i,j,k,l}\triangle t) \leq 1$ being the temporal correlation factor. $\mathcal{J}_0$ is the zero-order Bessel function of first kind, $f_{D,i,j,k,l}$ is the Doppler spread from the l-th antenna that belongs to the j-th cluster to k-th UE of the i-th cluster. Since the cooperative RRHs in one cluster are assumed static, we will assume that $\lambda_{i,j,k,1} = \lambda_{i,j,k,2} = \cdots = \lambda_{i,j,k,n_{\mathrm{A},j}} = \lambda_{i,j,k}$ in the rest of the paper for brevity. Finally, $\mathbf{e}_{i,j,k} \in \mathbb{C}^{1\times n_{\mathrm{A},j}}$ denotes the channel mismatch vector with each element being modeled as complex Gaussian distribution as $\mathcal{CN}(0, 1-\lambda^2_{i,j,k})$ [25].

Substituting (4) into (2) and considering (3), we obtain following relationship between the current and outdated channels as follows:

$$\begin{aligned}\bar{\mathbf{h}}_{i,j,k}[t] &= \lambda_{i,j,k}\mathbf{h}_{i,j,k}[t-\triangle t]\mathbf{G}_{i,j,k} + \mathbf{e}_{i,j}\mathbf{G}_{i,j,k} \\ &= \lambda_{i,j,k}\bar{\mathbf{h}}_{i,j,k}[t-\triangle t] + \mathbf{e}_{i,j}\mathbf{G}_{i,j,k}\ .\end{aligned} \tag{5}$$

It is apparent that since the calculation of the precoding matrix is based on the outdated channel $\bar{\mathbf{h}}_{i,j,k}[t-\triangle t]$ instead of the actual channel $\bar{\mathbf{h}}_{i,j,k}[t]$, the performance of the joint transmission process will be affected. Note that in the following, for simplification we define:

$$\hat{\mathbf{h}}_{i,j,k}[t] = \bar{\mathbf{h}}_{i,j,k}[t-\triangle t]\ , \tag{6}$$

and we omit the time index.

Thus, the received signal of the k-th UE in the i-th cluster can be written as:

$$\begin{aligned}y_{i,k} &= \underbrace{\sqrt{\gamma_i}\bar{\mathbf{h}}_{i,i,k}\mathbf{w}_{i,k}x_{i,k}}_{\text{Desired signal}} + \underbrace{\sqrt{\gamma_i}\sum_{l=1,l\neq k}^{n_{\mathrm{UE},i}}\bar{\mathbf{h}}_{i,i,k}\mathbf{w}_{i,l}x_{i,l}}_{\text{Intra-cluster interference}} \\ &+ \underbrace{\sum_{j=1,j\neq i}^{N_c}\sqrt{\gamma_j}\sum_{l=1}^{n_{\mathrm{UE},i}}\bar{\mathbf{h}}_{i,j,k}\mathbf{w}_{j,l}x_{j,l}}_{\text{Inter-cluster interference}} + \underbrace{n_{i,k}}_{\text{Noise}}\ ,\end{aligned} \tag{7}$$

where $n_{i,k}$ is the Gaussian noise with distribution $\mathcal{CN}(0,\sigma^2_{i,k})$, and $\mathbf{w}_{i,k} \in \mathbb{C}^{n_{\mathrm{A},i}\times 1}$ and $x_{i,k}$ stand for the precoding vector and transmit signal for the k-th UE in the i-th cluster, respectively. $x_{i,k}$ and $x_{j,l}$ are assumed uncorrelated for $(i,j)\neq(k,l)$ (for $i,j=1,\cdots,N_c$ and $k,l=1,\cdots,n_{\mathrm{U},i}$) and being zero mean and unit power variables, i.e. $\mathcal{E}\{\|x_{i,j}\|^2\}=1$. Finally, γ_i denotes the transmitting power of the signal $x_{i,k}$. Note that here we have supposed power is evenly allocated to each UE in a cooperative cluster.

By substituting (5) into (7), the desired signal power, P_x, and the interference power, P_I, for the k-th UE in the i-th cluster can be expressed as:

$$\begin{aligned}P_x &= \gamma_i\|\lambda_{i,i,j}\mathbf{h}_{i,i,k}\mathbf{G}_{i,i,k}\mathbf{w}_{i,k}x_{i,k}\|^2 + \gamma_i\|\mathbf{e}_{i,i,k}\mathbf{G}_{i,i,k}\mathbf{w}_{i,k}x_{i,k}\|^2 \\ P_I &= \|\sqrt{\gamma}_i\sum_{l=1,i\neq k}^{n_{\mathrm{UE},i}}(\lambda_{i,i,j}\mathbf{h}_{i,i,k}\mathbf{G}_{i,i,k} + \mathbf{e}_{i,i,k}\mathbf{G}_{i,i,k})\mathbf{w}_{i,l}x_{i,l}\|^2 \\ &+ \|\sqrt{\gamma}_j\sum_{j=1,j\neq i}^{N_c}\sum_{l=1}^{n_{\mathrm{UE},i}}\bar{\mathbf{h}}_{i,j,k}\mathbf{w}_{j,l}x_{j,l}\|^2\ ,\end{aligned} \tag{8}$$

where we have used the fact that $\mathbf{e}_{i,i,k}$ is independent of the channel vector $\mathbf{h}_{i,i,k}$ and $\mathbf{h}_{i,i,k}$ is independent of $\mathbf{h}_{i,j,k}$ when $i\neq j$. Thus, the expectation of the output SINR in the presence of delay impact can be written as:

$$\overline{\mathrm{SINR}}_{i,k} = \mathcal{E}\{\frac{P_x}{P_I+\sigma^2_{i,k}}\}\ . \tag{9}$$

III. Desired signal and Interference Power

The SINR expression in (9) is very difficult to analyze theoretically since it is a compound function of multiple variables including large-scale and fast fading of the channels, delay and precoding coefficients as well as multiple random deployed RRHs and users. As a solution, we will first derive the desired signal and interference power for single user case without considering delay impact and any specific precoding algorithm. In other words, we focus on user k and set $\mathbf{e}_{i,i,k} = \mathbf{0}$ and $\mathbf{w}_{i,l} = \mathbf{1}$, $\forall i,l$, in equation (9). However, we will consider all of these factors in the next section and we will demonstrate that the fast fading, and specific precoding coefficients along with delay-caused error can be treated independently.

Expressing the precoding fast fading and path-loss coefficients as functions of RRHs' location $z\in v$, the expected total received signal power at the k-th UE in the i-th cluster can be given as:

$$\begin{aligned}\bar{P}_x &= \mathcal{E}\{\sum_{z\in v}\|h(z)\|^2\|g(z)\|^2\|w(z)\|^2\|x\|^2\} \\ &= \rho_{\mathrm{RRH}}\int_{\mathbb{R}^d}\|g(z)\|^2dz\ .\end{aligned} \tag{10}$$

since $\mathcal{E}\{\|h(z)\|^2\}=\mathcal{E}\{\|w(z)\|^2\}=\mathcal{E}\{\|x\|^2\}=1$. When the UE of interest is assumed to be located at the *center* of the cluster, the received signal power can be given by [18]:

$$\begin{aligned}\bar{P}_x &= \rho_{\mathrm{RRH}}c_d d\int_0^R g(r)r^{d-1}dr \\ &= 2\pi\rho_{\mathrm{RRH}}\int_0^R r^{-\eta}rdr \qquad (\text{when}\quad d=2)\ .\end{aligned} \tag{11}$$

Note that in this work we consider that path-loss coefficients, $g(r)$, are derived from the following model:

$$g(r) = \begin{cases} R_0^{-\eta} & \text{if} \quad r\leq R_0 \\ r^{-\eta} & \text{if} \quad r > R_0 \end{cases}\ , \tag{12}$$

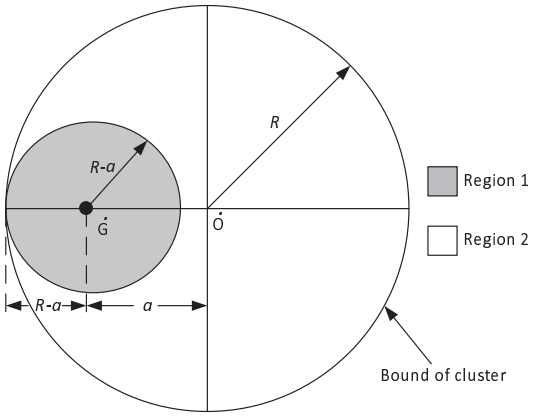

Fig. 3. The two integral regions of desired signal power for the UE at $\dot{G}$.

where η is the path-loss exponent[4] and R_0 is a minimum distance between the UE and RRH to bound the path-loss [18], i.e. the path-loss during the distance $[0, R_0]$ is assumed constant[5].

Equation (11) gives a general calculation method of the desired signal power for a UE located at the center of the network. In essence, this stands for the best case scenario on average since the UE receives the largest power from the distributed RRHs and smallest interference from outside of the cluster. However, the UE could be located at *arbitrary* point in the cluster, i.e. the location of UE is another random variable.

To solve the above problem we will take a step-by-step approach. Firstly, the desired signal and interference power will be derived considering the UE located at an *arbitrary* but *fixed* point in a cluster. Then, by treating the UE location as a random variable, the generic expressions will be derived. Note that all derivations will be for the more complex planar deployment in the first instance and then respective expressions for the linear case (e.g., for train line scenarios) will be provided.

A. UE at arbitrary fixed location

Without loss of generality, we assume that the UE is located at point $\dot{G}$ with a distance of $a \in [0, R]$ from the center of the cluster $\dot{O}$, as depicted in Fig. 3. Let's first consider the desired signal power, which is contributed by two regions (see Fig. 3), i.e.

$$\bar{P}_x(a) = \bar{P}_{x1}(a) + \bar{P}_{x2}(a) \ . \tag{13}$$

The first part of equation (13), $\bar{P}_{x1}(a)$, is attributed to the RRHs in region 1, i.e. within circular area $(\dot{G}, R-a)$, and it can be calculated straightforwardly by using equation (11) since this is a centrically symmetric region. Thus, replacing R with $R-a$ in (11), we get:

$$\bar{P}_{x1}(a) \ = \ 2\pi M \rho_{\mathrm{R}} \int_0^{R-a} g(r) r dr \ . \tag{14}$$

Depending on the values of $R-a$ and R_0 and considering (12), (14) could be expressed as:

$$\begin{aligned}\bar{P}_{x1}(a) &= \begin{cases} 2\pi M \rho_{\mathrm{R}} (\int_0^{R_0} R_0^{-\eta} r dr + \int_{R_0}^{R-a} r^{-\eta} r dr) \\ 2\pi M \rho_{\mathrm{R}} \int_0^{R-a} R_0^{-\eta} r dr \end{cases} \\ &= \begin{cases} 2\pi M \rho_{\mathrm{R}} (\frac{R_0^{2-\eta}}{2} + \frac{(R-a)^{2-\eta} - R_0^{2-\eta}}{2-\eta}) & \text{if } R-a > R_0 \\ 2\pi M \rho_{\mathrm{R}} \frac{(R-a)^2}{2} R_0^{-\eta} & \text{if } R-a \le R_0 \end{cases}\end{aligned} \tag{15}$$

The second part of equation (13), $\bar{P}_{x,2}(a)$, is attributed to the RRHs in region 2, i.e. within the subtraction area between circles $(\dot{O}, R)$ and $(\dot{G}, R-a)$. The calculation of this part is not straightforward due to its asymmetric shape and an integral method has to be used, as shown in Fig. 10 in Appendix A; for a fixed a, region 2 is divided into slim arcs with a length of $r\theta$ and depth of dr, where θ is the intersection angle between the circle $(\dot{O}, R)$ and $(\dot{G}, r)$ when $r > R-a$. By moving r from $R-a$ to $R+a$ and summing the power contribution from all slim arc areas, $\bar{P}_{x2}$ can be given by:

$$\bar{P}_{x2}(a) \ = \ M \rho_{\mathrm{R}} \int_{R-a}^{R+a} g(r) r \theta dr \ . \tag{16}$$

Finally, equation (16) can be written in terms of a as follows:

$$\bar{P}_{x2}(a) = 2M \rho_{\mathrm{R}} \int_{R-a}^{R+a} g(r) r \arccos \frac{r^2 + a^2 - R^2}{2ar} dr \ . \tag{17}$$

Proof: See Appendix A.

Unfortunately, there is no close-form solution for (17) when η takes most of the possible values. For that reason numerical methods are adopted to verify the effectiveness of the derivation in section V.

B. UE location as a random variable

So far, we have given the signal power of the UE with distance of a from the original point of the cluster circle $(\dot{O}, R)$. Next, we will treat a as a *random variable* which is evenly distributed in the cluster $(\dot{O}, R)$. The probability density function of a can be expressed as:

$$\varphi(a) = \frac{2a}{R^2} \quad 0 \le a \le R \ . \tag{18}$$

By considering equations (13), (15), (17) and (18), the desired signal within the cluster radius of R can be expressed as

$$\bar{P}_x = \bar{P}_{x1} + \bar{P}_{x2} = \int_0^R \frac{2a}{R^2} [\bar{P}_{x1}(a) + \bar{P}_{x2}(a)] da \ . \tag{19}$$

[4]Typical path loss exponent values vary between $2 \sim 3.5$ depending on the deployment scenarios, where the microcellular models suggest a smaller value of $\eta = 2$, and the macrocellular models suggest a much larger value of $\eta = 3.5$ [13].

[5]In some other channel path-loss modeling, an exclusive zero around the RRH may be introduced to bound the path-loss [17] or a uniform closed-form expression is used [13].

Using $\bar{P}_{x1}(a)$ from (15), the first part of (19) can be derived as:

$$\begin{aligned}\bar{P}_{x1} &= \int_0^R \frac{2a}{R^2}\bar{P}_{x1}(a)da \\ &= \int_0^{R-R_0} \frac{2a}{R^2}\bar{P}_{x1}(a)da + \int_{R-R_0}^{R} \frac{2a}{R^2}\bar{P}_{x1}(a)da \\ &= \frac{\pi\rho(R-R_0)^2R_0^2}{R^2}(1-\frac{2R^{-\eta}}{2-\eta}) + \frac{2\pi\rho[R_0^{3-\eta}-R^{3-\eta}]}{R(2-\eta)(3-\eta)} + \\ &\frac{2\pi\rho(-R_0)^{4-\eta}-(R)^{4-\eta}}{R^2(2-\eta)(4-\eta)}\pi\rho R_0^2[1-\frac{R-R_0^2}{R^2}] \,. \end{aligned} \quad (20)$$

Again, there is no closed-form solution for $\bar{P}_{x2}$, however, it can be calculated numerically by equation (16).

C. Interference power

Unlike the desired signal power, the expression of the interference power cannot be derived straightforwardly; as can be seen from Fig.2, the integral region is irregular and also depends on the location of the cluster within the network. However, it can be obtained indirectly by deriving the received power from the whole network and subtracting the desired signal power part. Considering the UE located at the center of the network, the total power received by the UE can be easily obtained using equation (11):

$$\begin{aligned}\bar{P}_{tot} &\approx 2\pi\rho_{\text{RRH}}\int_0^{R_t} g(r)rdr \\ &= \pi\rho_{\text{RRH}}R_0^2 + 2\pi\rho_{\text{RRH}}\frac{R_t^2-R_0^2}{2-\eta} \,. \end{aligned} \quad (21)$$

When the UE is at a fixed point in the cluster with a distance of a from the center of the cluster, the interference power contributed by the area outside of the cluster is:

$$\bar{P}_I(a) = \bar{P}_{tot} - \bar{P}_x(a) \,. \quad (22)$$

Similarly, for a UE at a random location inside the cluster, the interference power can be expressed as

$$\bar{P}_I = \bar{P}_{tot} - \bar{P}_x \,. \quad (23)$$

Equations (22) and (23) are very accurate approximations when the cluster is in the center of the network and the interfering area is much larger than the cluster size. In most cases, it is a practical assumption since the defined network with limited radius R_t will be surrounded by other networks and hence receive interference from them. We will see in the results section V that this approximation is quite accurate even for small number of clusters in the network.

D. Linear deployment

For the linear deployment, cluster size R refers to length from the center to the edge of the cluster. Since $d = 1$ and $c_d = 2$ in that case, equation (11) becomes:

$$\bar{P}_x = 2M\rho_{\text{R}}\int_0^R g(r)dr \qquad (\text{when} \quad d=1) \,. \quad (24)$$

Following the same derivation as for the planar case, we can obtain the desired signal power based on an arbitrary UE location in a closed-form expression as:

$$\bar{P}_x^L(a)=\begin{cases} M\rho_{\text{R}}\left[(R_0+R-a)R_0^{-\eta}+\frac{(R+a)^{1-\eta}-R_0^{1-\eta}}{(1-\eta)}\right] & \text{if} \quad R-a\le R_0 \\ M\rho_{\text{R}}\left[2R_0^{1-\eta}+\frac{(R+a)^{1-\eta}+(R-a)^{1-\eta}-2R_0^{1-\eta}}{(1-\eta)}\right] & \text{if} \quad R-a>R_0 \end{cases} \quad (25)$$

where the superscript $\{\cdot\}^L$ is used to differentiate from the planar case. By treating the location parameter a as a random variable with even distribution on $[0, R]$, the desired signal power can be given by:

$$\bar{P}_x^L = \int_0^R \frac{1}{R}\bar{P}_x^L(a)da \,. \quad (26)$$

Substituting (25) into (26), we have the *closed-form expression* for the desired signal power for a randomly positioned UE:

$$\begin{aligned}\bar{P}_x^L &= \frac{\rho_{\text{RRH}}}{R}[(2RR_0^{1-\eta}-\frac{R_0^{2-\eta}}{2}) + \frac{(2R)^{2-\eta}-R_0^{2-\eta}}{(1-\eta)(2-\eta)} \\ &+ \frac{2R(R_0)^{1-\eta}-R_0^{2-\eta}}{(1-\eta)}] \,. \end{aligned} \quad (27)$$

The total signal power received can be expressed as $\bar{P}_{tot}^L = M\rho_{\text{R}}[2R_0^{1-\eta}+2\frac{R_t^{1-\eta}-R_0^{1-\eta}}{1-\eta}]$, thus, the interference power can be approximately obtained by subtracting the desired signal power from the total power:

$$\begin{aligned}\bar{P}_I^L &\approx \bar{P}_{tot}^L - \bar{P}_x^L = \frac{\rho_{\text{RRH}}}{R}[\frac{R_0^{2-\eta}-(2R)^{2-\eta}}{(1-\eta)(2-\eta)} \\ &+\frac{2RR_t^{1-\eta}-2RR_0^{1-\eta}+R_0^{2-\eta}-2RR_0^{1-\eta}}{(1-\eta)}] - \frac{R_0^{2-\eta}}{2} \,. \end{aligned} \quad (28)$$

So far we have proposed a generic approach for evaluating the desired signal and interference power at UEs in randomly deployed small cell networks. In the next, we will focus on the optimization of cluster size in order to maximize system performance.

IV. Cluster Size Optimization

In this section, we first derive the expectation of the SINR in the presence of small-scale fast fading and the delay for two representative precoders: MRT and ZF, respectively. The CSI latency model in terms of cluster size and cloud/RRH configuration will be built in Section IV-B. Then, the optimization problem in the criterion of maximizing ergodic sum-rate in terms of cluster size will be formulated in Section IV-C.

A. Output SINR

Let's define the precoding matrix and observed channel matrix for the i-th cluster as $\mathbf{W}_i = [\mathbf{w}_{i,1}, \mathbf{w}_{i,2}, \cdots, \mathbf{w}_{i,n_{\text{U},i}}]$ and $\hat{\mathbf{H}}_i = [\hat{\mathbf{h}}_{i,i,1}; \hat{\mathbf{h}}_{i,i,2}; \cdots; \hat{\mathbf{h}}_{i,i,n_{\text{A},i}}]$, respectively. Then the MRT and ZF precoders can be expressed as [11]:

$$\begin{aligned}\mathbf{w}_{i,k}^{MRT} &= \frac{\hat{\mathbf{h}}_{i,i,k}^H}{\|\hat{\mathbf{h}}_{i,i,k}\|} \\ \mathbf{w}_{i,k}^{ZF} &= \frac{\hat{\mathbf{H}}_i^H(\hat{\mathbf{H}}_i\hat{\mathbf{H}}_i^H)^{-1}\mathbf{1}_k}{\|\hat{\mathbf{H}}_i^H(\hat{\mathbf{H}}_i\hat{\mathbf{H}}_i^H)^{-1}\mathbf{1}_k\|} \,, \end{aligned} \quad (29)$$

where $\mathbf{1}_k = [0, \cdots, 0, 1, 0, \cdots]^T$ refers to a vector with its k-th element being 1 and all other elements being 0. Thus, $\hat{\mathbf{H}}_i^H(\hat{\mathbf{H}}_i\hat{\mathbf{H}}_i^H)^{-1}\mathbf{1}_k$ refers to the k-th column of matrix $\hat{\mathbf{H}}_i^H(\hat{\mathbf{H}}_i\hat{\mathbf{H}}_i^H)^{-1}$.

1) Output SINR with MRT precoder: By substituting the first equation of (29) into (9) and considering the random vector $\mathbf{e}_{i,j,k}$ independent of $\mathbf{h}_{i,j,k}$, along with assumption that each cluster consumes the same power for transmission, i.e. $\gamma_1 = \gamma_2 = \cdots = \gamma_{N_c}$, we get:

$$\overline{\text{SINR}}_{i,k}^{MRT} \approx [\lambda_{i,i,k}^2 + \frac{1-\lambda_{i,i,k}^2}{MN_{\text{R}}}]\bar{P}_x \cdot P_{I,exp}^{MRT} , \quad (30)$$

where

$$P_{I,exp}^{MRT} = \sum_{l=1}^{N_{\text{U}}-2}(-1)^{N_{\text{U}}-2-l}\sum_{n=0}^{l-1}\frac{P_{I,2}^{N_{\text{U}}-2-l+n}\xi_1^{N_{\text{U}}-1+n-l}}{l(N_{\text{U}}-2-l)n!} + \int_0^{\infty} e^{P_{I,2}\xi_1}(-P_{I,2})^{N_{\text{U}}-2}\frac{e^{(P_{I,2}+u)\xi_1}}{(u+P_{I,2})(N_{\text{U}}-2)!}du \quad (31)$$

with $\xi_1 = \frac{MN_{\text{R}}}{\gamma_1 P_x}$ and $P_{I,2} = \gamma_1\frac{N_{\text{U}}}{MN_{\text{R}}}\bar{P}_I + \sigma_{i,k}^2$

Proof: See appendix B.

2) Output SINR with ZF precoder: Similarly, considering ZF precoder the output SINR in the presence of latency can be expressed as:

$$\overline{\text{SINR}}_{i,k}^{ZF} \approx [\lambda_{i,i,k}^2 + \frac{1-N_{\text{U}}\lambda_{i,i,k}^2}{MN_{\text{R}}}]\bar{P}_x \cdot P_{I,exp}^{ZF} , \quad (32)$$

where $P_{I,exp}^{ZF}$ has the same expression as $P_{I,exp}^{MRT}$ except replacing ξ_1 by ξ_2 in equation (31) with $\xi_2 = \frac{MN_{\text{R}}}{\gamma_1(1-\lambda_{i,i,1}^2)P_x}$.

Proof: See appendix C.

Comparing the first part (before the multiplication sign) of (30) and (32), we observe that, when $N_{\text{U}} > 1$, MTR-based desired signal power will be always larger than the ZF-based one. This is due to the fact that ZF uses the spatial (i.e. antenna) degrees of freedom (DoF) to eliminate interference while MRT explores all DoF to maximize the desired signal power. Comparing the second part (after the multiplication sign) of (30) and (32), the only difference is that an extra term $(1-\lambda_{i,i,1}^2)$ is multiplied with ξ_1 in the case of ZF precoding. This term is essentially the residual intra-cluster interference due to the delay caused mismatch; larger delay leads to smaller $\lambda_{i,i,1}^2$ and higher the residual intra-cluster interference. Note that in case of MRT precoding, this second part is not affected by the introduced delay.

In order to express the output SINR as a function of the cluster size R, taking planar case as an example, we can substitute $v = 2\pi R^2$, $V = 2\pi R_t^2$, $\lambda = \mathcal{J}_0(2\pi f_D \Delta t)$ and equation (1) into equations (30) and (32). Then the output SINR for the random distributed RRHs and UEs in the presence of delay for MRT precoding could be expressed as:

$$\overline{\text{SINR}}_{i,k}^{MRT} \approx [\lambda(R)^2 + \frac{(1-\lambda(R)^2)}{2\pi R^2 M\rho_{\text{R}}}]\bar{P}_x \cdot P_{I,exp}^{MRT}(R) , \quad (33)$$

and for ZF precoding:

$$\overline{\text{SINR}}_{i,k}^{ZF} \approx \left[\frac{[2\pi R^2(M\rho_{\text{R}}-\rho_{\text{U}})]\lambda(R)^2+1}{2\pi R^2 M\rho_{\text{R}}}\right]\bar{P}_x \cdot P_{I,exp}^{ZF}(R) , \quad (34)$$

where the temporal correlation factor λ is expressed as a function of R and its subscripts are omitted for brevity.

B. Delay model

In general, the total delay of the precoding process is caused by several factors such as the pilot estimation and processing delay at the UE, propagation delay from UE to RRH and from RRH to cloud, CSI feedback (and scheduling) delay, RRH processing delay, cloud data center processing delay and backhaul latency. Thus, total delay can be generally modelled as:

$$\begin{aligned}\Delta t &= \varpi_1\Delta t_{chan-est} + \varpi_2(\Delta t_{fb} + \Delta t_{prop-tot}) \\ &+ \Delta t_{process-cloud} + \Delta t_{process-RRH} + \varpi_3\Delta t_{BH} ,\end{aligned} \quad (35)$$

where

$$\begin{aligned}\varpi_1 &= MN_{\text{R}} = vM\rho_{\text{R}}; \\ \varpi_2 &= N_{\text{U}}MN_{\text{R}}N_c/q_{fb} = Vv\rho_{\text{U}}M\rho_{\text{R}}/q_{fb} .\end{aligned} \quad (36)$$

The physical meaning of each item in (35) is explained one-by-one in the following subsections.

1) Channel estimation delay: The first item in (35) $\Delta t_{chan-est}$, denotes the channel estimation delay at the UE and ϖ_1 stands for the number of channel coefficients to be estimated for one UE. Apparently, the more channels to be estimated, the larger the delay is likely to be.

2) CSI feedback and propagation delay: Δt_{fb} and $\Delta t_{prop-tot}$ in (35) denote the average per channel coefficient feedback delay and total propagation delay, respectively. $N_{\text{U}}MN_{\text{R}}N_c$ stands for the total number of channel coefficients and q_{fb} is a factor denoting how many channels can be fed back each time. ϖ_2 stands for the total number of times CSI is to be fed back for the whole network. Assuming that CSI feedback from UE to RRH has a capacity of C_{fb} and considering that each CSI is quantized to B_1 bits, the feedback delay can be written as $\Delta t_{fb} = B_1/q_{fb}$. Moreover, the total propagation delay can be expressed as $\Delta t_{prop-tot} = 2(s_{u2r} + s_{r2c})/cq_{fb}$, where c is the speed of light, and s_{u2r} and s_{r2c} are the distances from UE to RRH and from RRH to cloud, respectively.

3) Cloud processing delay: $\Delta t_{process-cloud}$ in (35) denotes the cloud processing delay, which is composed of two factors and can be written as $\Delta t_{process-cloud} = \Delta t_{Tx1} + \Delta t_{precoder-cal}$. Δt_{Tx1} is attributed to the (part of) baseband processing (such as coding, modulation, precoding, IFFT, etc.) depending on the transmission chain functionality split between cloud and RRH [1], [2]. In general, the total delay-caused by baseband processing at the transmitter[6] $\Delta t_{Tx} = \Delta t_{Tx1} + \Delta t_{Tx2}$ is assumed to be constant, where Δt_{Tx2} refers to the respective delay at RRH. $\Delta t_{precoder-cal}$ stands for the precoder calculation delay, which is a dominating factor when the cluster size is relatively large and the available computational resource is limited. Note also that different precoding algorithms lead to dramatically different computational complexity. For example, ZF has significant larger complexity

[6]In the current LTE-A protocol, the total Tx processing time left to eNB and UE is around 3ms including the propagation delay [26], [27]. The worst case of Tx processing time is around 2.3ms which corresponds to the case of cell radius being 100km and the propagation delay is 0.6 ms [26], [27], i.e. $\Delta t_{Tx} \in [2.3, 3]$ ms.

than MRT precoding. Taking ZF as an example[7], the delay caused by the precoding matrix calculation can be written as:

$$\Delta t_{precoder-cal} = \frac{(K_{add}^{ZF} + \zeta_2 K_{multi}^{ZF})V}{vC_{com}q_c} , \quad (37)$$

where K_{multi}^{ZF} and K_{add}^{ZF} denote the required real-time operations of multiplication and addition, respectively; ζ_2 is the equivalent addition operation times for each multiplication; C_{com} denotes the cloud computational capability; and q_c is the resource division factor (since the computational resources is likely shared by multi-tasks, and assuming a uniform distribution of resources, $1/q_c$ of the total available resources will be allocated to the precoding matrix calculation). Thus $C_{com}q_c$ available computational capability will be allocated in total to the precoding matrix calculation. Note also that q_c could be set smaller than 1 corresponding to the case where multiple processors could contribute to the computation in parallel. The values of K_{multi}^{ZF} and K_{add}^{ZF} depend on the number of users and RRHs in the cluster and can be calculated as:

$$\begin{aligned} K_{multi}^{ZF} &= 8N_{UE}^2 N_{\text{RRH}} + \mathcal{O}(4N_{\text{UE}}^3) + 2N_{\text{RRH}}N_{\text{UE}} \\ &= 8v^3\rho_{UE}^2\rho_{\text{RRH}} + O(4v^3\rho_{\text{UE}}^3) + 2v^2\rho_{\text{UE}}\rho_{\text{RRH}} \end{aligned} \quad (38)$$

$$\begin{aligned} K_{add}^{ZF} &= 8N_{UE}^2 N_{\text{RRH}} - 2N_{UE}^2 \\ &+ O(4N_{\text{UE}}^3) - 2N_{\text{UE}} + 2\zeta_1 N_{\text{RRH}}N_{\text{UE}} \\ &= 8v^3\rho_{UE}^2\rho_{\text{RRH}} - 2v^2\rho_{\text{UE}} + O(4v^3\rho_{\text{UE}}^3) + 2v^2\rho_{\text{UE}}\rho_{\text{RRH}} , \end{aligned} \quad (39)$$

where the term $\mathcal{O}(4N_{\text{U}}^3)$ arises from the matrix inversion process and its complexity depends on the specific implemented algorithm (its typical value takes $8/3$ [28]). Moreover, ζ_1 factor indicates how much time the multiplication process consumes compared to the addition process.

4) RRH processing delay: $\Delta t_{process-RRH}$ in (35) denotes the RRH processing delay which also comprises of two parts, i.e. $\Delta t_{process-RRH} = \Delta t_{Tx2} + \Delta t_{CSI-fw}$. The first item, Δt_{Tx2}, as already mentioned before is attributed to (part) of the baseband and RF implementation at the RRH. The second item, Δt_{CSI-fw}, stands for the delay due to CSI feedback from RRHs to the cloud in the uplink. Apparently, the total processing delay (both at the cloud and RRH) can be given by $\Delta t_{Tx} + \Delta t_{precoder-cal} + \Delta t_{CSI-fw}$.

5) Backhaul latency: The last item contains the backhaul latency given by the backhaul latency Δt_{BH} per hop multiplied with the number of backhaul hops ϖ_3 from the cloud data center to the small cells. The values for Δt_{BH} of various backhaul technologies can be found in (Section 6.3 of) [29] and [30], [31].

C. Ergodic sum-rate and optimization formulation

The optimization problem for maximizing the ergodic sum-rate of the network in terms of cluster size can be expressed as:

$$\max_R C_s \qquad \text{subject to} \qquad 0 < R \leq R_t , \quad (40)$$

[7]Due to the factor that MRT precoding algorithm has very low computational complexity, the precoding calculation caused delay is negligible and omitted here.

where ergodic sum-rate R_s in our system model can be given by:

$$\begin{aligned} R_s &= \mathcal{E}\{\sum_{i=1}^{N_c}\sum_{k=1}^{n_{\text{UE},i}} log_2(1+SINR_{i,k})\} \\ &= N_c\mathcal{E}\{\sum_{k=1}^{n_{\text{UE},i}} log_2(1+SINR_{i,k})\} \\ &= \rho_{\text{UE}}V\mathcal{E}\{log_2(1+SINR_{i,k})\} \\ &\leq \rho_{\text{UE}}Vlog_2(1+\mathcal{E}\{SINR_{i,k}\}) . \end{aligned} \quad (41)$$

where $SINR_{i,k}$ is instantaneous SINR of the k-th UE at the i-th cluster. The second equation of (41) is based on the assumption that the cellular network is surrounded by other non-overlapping and same configured networks. In that case, each cluster can be effectively considered at the center of a network, therefore, equally contributing to the sum-rate. The third equation in (41) holds when each UE is randomly and independently distributed within the network, where the total number of the UEs can be given in terms of the UE density and network volume as $\mathcal{E}\{N_cN_{\text{U}}\} = \rho_{\text{U}}V$. Since it is very difficult to solve the ergodic sum-rate directly due to the expectation operation implemented outside of the logarithm, the well-known Jensen's inequality is used in the fourth step of (41) to obtain an upper bound. Although the ergodic sum-rate upper bound can be a loose bound in some cases, in results section V we show that the optimal cluster size obtained from the proposed analytical framework match very well with the simulation results.

Thus, by considering MRT precoding and writing $\mathcal{E}\{SINR_{i,k}\} = \overline{\text{SINR}}_{i,k}^{MRT}$, the optimization problem in (40) is approximately equivalent to:

$$\max_R \overline{\text{SINR}}_{i,k}^{MRT} \qquad \text{subject to} \qquad 0 < R \leq R_t , \quad (42)$$

and similarly for ZF precoding we have:

$$\max_R \overline{\text{SINR}}_{i,k}^{ZF} \qquad \text{subject to} \qquad 0 < R \leq R_t . \quad (43)$$

Note that, since the cost function is complex in terms of R, it is difficult to obtain the closed-form optimal solution and in section V, the numerical methods are adopted to verify its effectiveness.

V. Simulation results

In this section, we use Monte-Carlo simulations to investigate the proposed clustering optimization problem with ZF and MRT precoding algorithms for linear and planar dense small cell deployments. 1000 small cells and 1000 active UEs are uniformly distributed in a) a circular network area (planar deployment) with radius $R_t = 500$ meters and b) a linear network segment (linear deployment) of length $R_t = 1000$ meters. We assume the number of antennas at each RRH $M = 2$. For planar deployment, in order to approximate a circle bounded network we consider clusters formed by 1 to 7 tiers of cells, i.e. by considering that tier-1 consist of 7 cell, tier-2 of 19 cells and so on, cluster size will take values from the range set [7, 19, 37, 61, 91, 127, 169]. The input signal-to-noise (SNR) power is set to 30dB. The path-loss exponent

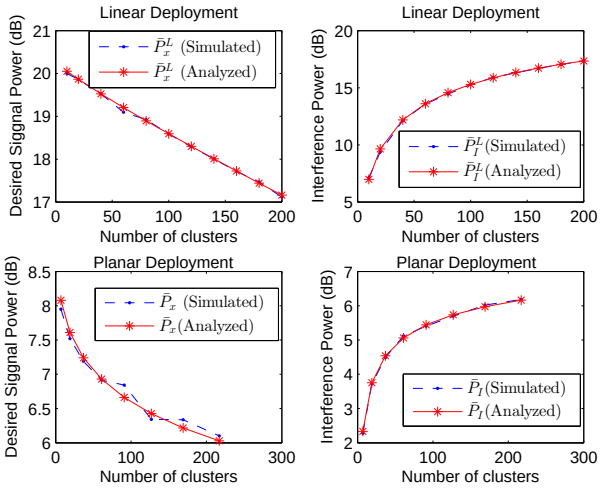

Fig. 4. Desired signal and interference power for planar and linear small cell deployments in the absence of delay.

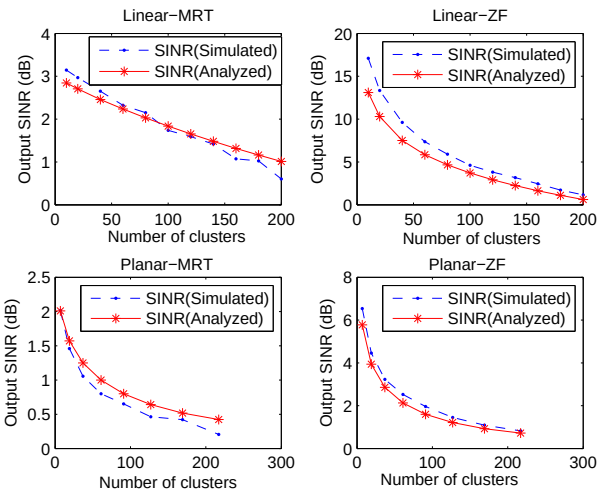

Fig. 5. Output SINR with ZF and MRT precoding for planar and linear small cell deployments in the absence of delay.

is kept $\eta = 2.2$ and Rayleigh fast fading is considered to model channels between RRHs and UEs, as given by equation (3). The temporal correlation of the channel is modeled by equation (4) with Doppler spread $f_D = 10$ Hz for all links. In either linear and planar deployment, $R_0 = 5$ meters. Without loss of generality, we only consider the performance of the UEs in the central cluster and assume interference from outside of the network to these UEs is negligible.

1) $\bar{P}_x$, $\bar{P}_I$ and output SINR in the absence of latency: To investigate the effectiveness of our theoretical analysis (i.e. proposed system, channel and clustering model in Sections III and IV), we first evaluate $\bar{P}_x$, $\bar{P}_I$ and output SINR in the absence of latency, i.e., $\Delta t = 0$ and $\lambda^2 = 1$.

$\bar{P}_x$ and $\bar{P}_I$ are evaluated using equation (19) and (23) for planar deployment (equations (27) and (28) for linear deployment) and compared with simulation results in Fig. 4, for different number of clusters within the network. It can be seen that the analytical results for both the desired signal and interference power match the simulation results perfectly for both planar and linear deployments. As expected, the desired signal power reduces with cluster size (i.e. larger number of clusters in the network) since less number of cooperating small cells contribute to the desired power. On the other hand, the interference power becomes larger as the number of clusters increases since the total number of interfering RRH outside the cluster is increased.

Furthermore, the output SINR is evaluated using equation (30) and (32) for MRT and ZF precoding respectively and compared with simulation results in Fig. 5. We observe that all four graphs (Linear-MRT, Linear-ZF, Planar-MRT and Planar-ZF) show good consistency between analytical and simulation results. For all cases, the output SINR decreases with increasing number of clusters due to the fact that a smaller cluster tends to obtain less desired signal power while more interference is caused from outside the cluster. Note that small gaps can be observed between theoretical and simulated results due to the assumptions used in clustering model.

In the following we only focus on the more complex planar deployment due to similar observed behavior of planar and linear case in terms of output SINR. Nevertheless, analogous results can be provided for the linear case as well, following the respective analytical expressions and simulation model.

2) Performance optimization in the presence of latency: To evaluate performance in the presence of latency, we first set some practical values for the parameters in the delay model. To this end, we consider the worst-case of Tx processing time defined in 3GPP, i.e. $\Delta t_{Tx} = 2.3$ ms [26], [27]; the channel estimation and RRH processing delay are set to zero since they are not as sizeable factors as the processing or feedback delay. The factors for multiplication and division over addition are set to the typical values of $\zeta_1 = 1$ and $\zeta_2 = 10$, respectively. The average distance between UE and RRH plus the distance between RRH and cloud data center is considered to be $s_{u2r} + s_{r2c} = 1000$ meters. The cloud is assumed to contain an Intel Xeon Processor E5-2680 with processing capacity of 1.73×10^{11} double-precision floating-point operations per second (DP-FLOPS)[8]. Assuming the addition operation is double-precision floating (64-bit), the available processing capacity for precoding calculation becomes $C_{com} = 1.73 \times 10^{11}/q_c$ addition operations per second. Furthermore, the computational resource division factor q_c is set to be 10, unless specified otherwise. The feedback capacity of the backhaul link from RRHs to cloud is set to $C_{fb} = 10^7$ bits per second and, unless specified otherwise, we set $q_{fb} = 1$, i.e. feedback of one channel coefficient each time. We assume that each small cell is connected to the cloud data center by one hop Dark fibre with latency $\Delta t_{FB} = 10$ micro second per kilometer per hop as defined by [29].

Fig. 6 and Fig. 7 investigate the delay impact on output SINR and sum-rate of the ZF-based algorithm in terms of

[8]8-core Intel Xeon Processor E5-2680 has a CPU frequency 2.7×10^9 and 2 operations per clock period, supporting 256-bit Advanced Vector Extensions (AVE), therefore we can calculate the computational capability as $8 \times 2 \times 2.7 \times 10^9 \times 256/64 = 1.73 \times 10^{11}$ DP-FLOPS.

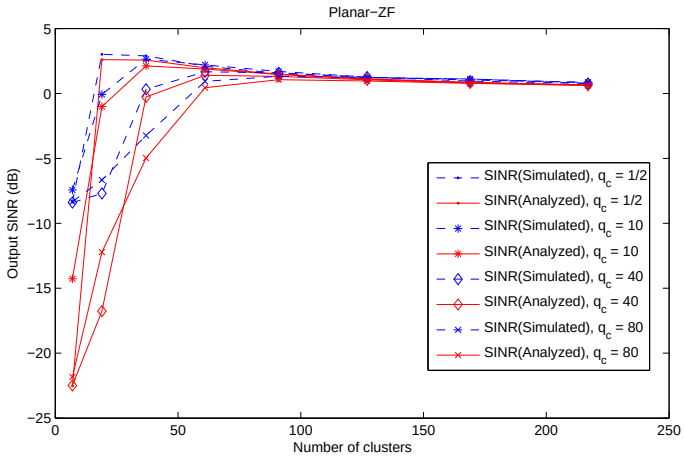

Fig. 6. Output SINR in the presence of delay for ZF-based planar deployment under various available computational resource division factors q_c.

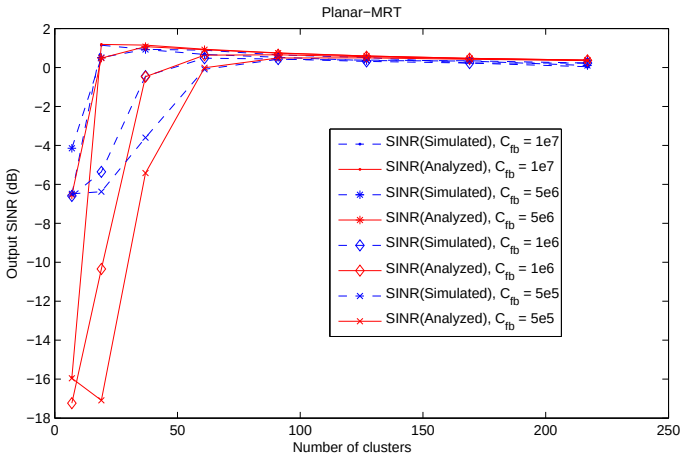

Fig. 8. Output SINR in the presence of delay for MRT-based planar deployment for different feedback capacity C_{fb}.

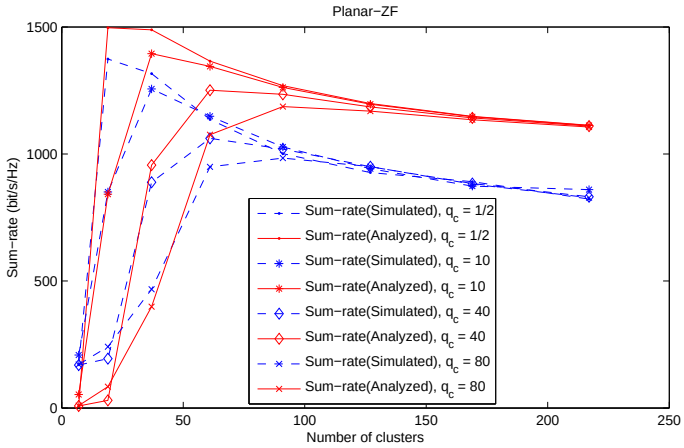

Fig. 7. Sum-rate in the presence of delay for ZF-based planar deployment under various available computational resource division factors q_c.

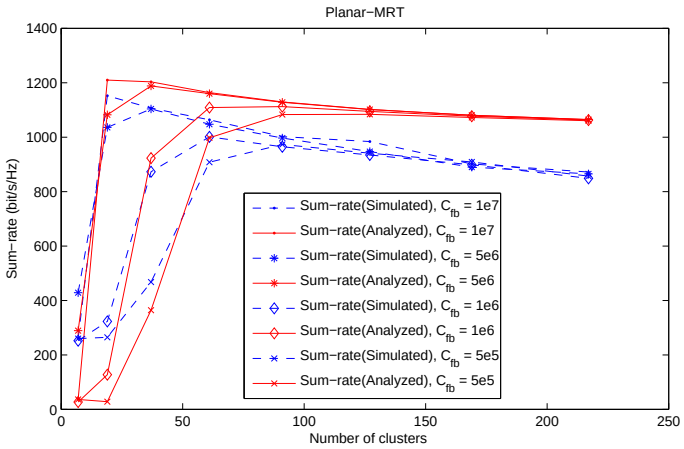

Fig. 9. Sum-rate in the presence of delay for MRT-based planar deployment under different feedback capacity C_{fb}.

cluster size. Various cases of cloud processing capability are considered with the computational load changing from 2, 1/10, 1/40 to 1/80 (i.e. q_c changing from 1/2, 10, 40, to 80). We observe that the analytical results (i.e. using equations (30) and (32)) roughly match the corresponding simulation results, especially for output SINR. More importantly, the peak points denoting the optimal cluster size are strictly overlapping with each other, for any specific configuration; thus, the optimal cluster size evaluation is not affected by the approximation in equation (41) where Jensen's inequality and the upper bound of the sum-rate have been considered (obviously, the theoretic sum-rate for each q_c is generally higher than the simulated results). We also note that the optimal cluster size decreases (i.e. optimal number of clusters increases) as the computational capability factor becomes larger. This is due to the fact that when q_c increases, less computational resources become available, therefore, smaller cluster size is needed to keep the delay-caused channel mismatch at low levels.

Fig. 8 and Fig. 9 illustrate the impact of feedback delay on the output SINR and sum-rate for various cluster sizes under MRT-based precoding, where the precoding matrix calculation delay is negligible. Compared to the results shown in Fig. 6 and Fig. 7 regarding the processing delay, the curve slopes are more flat for increasing number of clusters. This is due to the fact that the feedback-caused latency is only in the second order of the number of cooperative antennas; on the other hand, latency caused by the processing capability is in the cubic order of the number of cooperative antennas, thus, decreasing cluster size will lead to faster reduction of latency.

VI. Conclusions

Based on commonly used linear precoding algorithms (ZF and MRT) and linear and planar small cell deployment configurations, the paper proposed an approach for cluster size optimization in cloud-based distributed cooperative small cell networks in the presence of CSI latency, which is mainly caused by cloud processing delay and CSI feedback delay. An optimization problem is formulated in the aforementioned framework and desired signal and interference signal are calculated, which is followed by derivation of the output SINR

by taking into consideration the channel mismatch caused by latency due to small cells cooperation. Both delay and output SINR have been derived as a function of cooperation cluster size and an optimization problem to trade off the interference and channel mismatch has been formulated for maximizing network sum-rate. Simulations reveal a small gap with the analytical results in terms of SINR and sum-rate evaluations, and the proposed concise analytical framework can be safely used in order identify the optimal cluster size for any specific deployment.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007 - 2013 under grant agreement no. 317941 - project iJOIN. The European Union and its agencies are not liable or otherwise responsible for the contents of this document; its content reflects the view of its authors only. We gratefully recognise the great contributions of many colleagues from iJOIN, who in fruitful cooperation, contributed with valuable insight, surveys and vision The author would also like to acknowledge the support of the University of Surrey 5GIC (http://www.surrey.ac.uk/5gic) members for this work.

APPENDIX A
PROOF OF EQUATION (17)

Depending on the values of a and r, the proof of (17) is divided into two cases as depicted in Fig. 10. An auxiliary dot line ($\dot{J}\dot{K}$) is drawn in Fig. 10 to split the integral region 2 (circle with radius R and center at $\dot{O}$) into two parts. The left-hand side corresponds to $r^2 \leq R^2 - a^2$, which implies when $\theta \geq \pi$ and the supplementary angles will be calculated (Fig. 10 (a)). The right-hand side of ($\dot{J}\dot{K}$) corresponds to $r^2 \geq R^2 - a^2$ and $\theta \leq \pi$ (Fig. 10 (b)). We solve both cases using different derivation procedures and we show that both lead to an unified expression of the angle θ.

Let's first consider the case where $r^2 \geq R^2 - a^2$. The angle θ can be expressed by:

$$\theta = 2\text{arccos}\frac{\dot{G}\dot{D}}{\dot{G}\dot{A}} = 2\text{arccos}\frac{\dot{G}\dot{D}}{r} \ , \tag{44}$$

where $\dot{A}\dot{D}$ is perpendicular to $\dot{F}\dot{C}$. Considering the rectangular triangle $\dot{A}\dot{B}\dot{C}$ inscribed in the circle $\dot{O}$, according to the projective theorem $\dot{A}\dot{D} = (\dot{B}\dot{D})\cdot(\dot{C}\dot{D}) = (R-a+\dot{G}\dot{D})(R+a-\dot{G}\dot{D})$ [37]; similarly, when we consider the rectangular triangle $\dot{A}\dot{F}\dot{E}$ inscribed in the circle with radius r and circle center $\dot{G}$, we have $\dot{A}\dot{D} = (\dot{F}\dot{D})\cdot(\dot{E}\dot{D}) = (r-\dot{G}\dot{D})(r+\dot{G}\dot{D})$, which leads to $\dot{G}\dot{D} = (r^2 + a^2 - R^2)/(2a)$. Substituting $\dot{G}\dot{D}$ into (44), we can obtain θ as follows:

$$\theta = 2\text{arccos}\frac{r^2 + a^2 - R^2}{2ar} \ . \tag{45}$$

In the case where $\theta \geq \pi$, we have:

$$\theta = 2\pi - 2\text{arccos}\beta = 2\pi - 2\text{arccos}\frac{\dot{G}\dot{D}}{r} \ . \tag{46}$$

Considering the rectangular triangles $\dot{A}\dot{B}\dot{C}$ and $\dot{A}\dot{F}\dot{E}$ and using the projective theorems, we obtain: $\dot{A}\dot{D} = (\dot{B}\dot{D})\cdot(\dot{C}\dot{D}) = (R-a-\dot{G}\dot{D})(R+a+\dot{G}\dot{D})$ and $\dot{A}\dot{D} = (\dot{F}\dot{D})\cdot(\dot{E}\dot{D}) = (r-\dot{G}\dot{D})(r+\dot{G}\dot{D})$, which leads to

$$(\dot{G}\dot{D}) = \frac{-r^2 - a^2 + R^2}{2a} \ . \tag{47}$$

Substituting (47) into (46), we obtain (45), which means that in both cases provide the same expression of θ. Thus, substituting (45) into (16), we obtain (17).

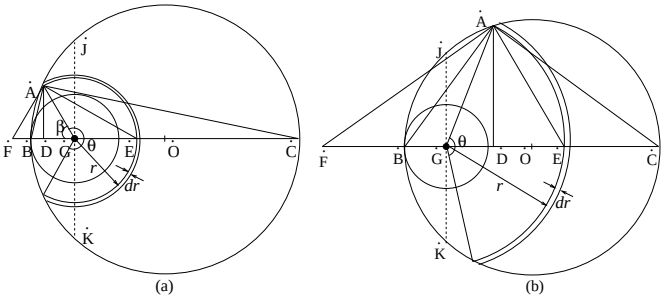

Fig. 10. Region 2 integral diagram for the desired signal power in the cluster. (a), case 1: $r^2 \leq R^2 - a^2$. (b), case 2: $r^2 \geq R^2 - a^2$.

APPENDIX B
PROOF OF EQUATION (30)

For the derivation of equation (9), we use the following approximation [32], [33]:

$$\overline{\text{SINR}}_{i,k}^{MRT} = \mathcal{E}\{\frac{P_x}{P_I + \sigma_{i,k}^2}\} \approx \quad \mathcal{E}\{P_x\}\mathcal{E}\{\tfrac{1}{P_I+\sigma_{i,k}^2}\} \ . \tag{48}$$

and we derive $\mathcal{E}\{P_x\}$ and $\mathcal{E}\{\frac{1}{P_I+\sigma_{i,k}^2}\}$ one-by-one. Let us first consider the desired signal power; substituting the first equation of (29) into the first equation of (8), the power of the desired signal can be expressed as:

$$P_x = \gamma_i \lambda_{i,i,k}^2 \|\mathbf{h}_{i,i,k}\mathbf{G}_{i,i,k}\|^2 + \gamma_i \frac{\|\mathbf{e}_{i,i,k}\mathbf{G}_{i,i,k}\mathbf{G}_{i,i,k}^H \mathbf{h}_{i,i,k}^H\|^2}{\|\mathbf{h}_{i,i,k}\mathbf{G}_{i,i,k}\|^2} \ . \tag{49}$$

Considering that each diagonal element in $\mathbf{G}$ matrices can be approximated by the average of all the diagonal elements:

$$g_{i,i,k,l}^2 \approx \frac{1}{n_{\text{A},i}} \sum_{m=1}^{n_{\text{A},i}} g_{i,i,k,m}^2 \quad \text{for} \quad l = 1, 2, \cdots, n_{\text{A},i} \tag{50}$$

the first item of (49) can be written as $\gamma_i \lambda_{i,i,k}^2 \frac{1}{n_{\text{A},i}} \sum_{m=1}^{n_{\text{A},i}} g_{i,i,k,m}^2 \sum_{l=1}^{n_{\text{A},i}} \|h_{i,i,k,l}\|^2$, where $h_{i,i,k,l}$ denotes the l-th element of the channel vector $\mathbf{h}_{i,i,k}$. Similarly, the second item of equation (49) can be simplified as $\frac{\gamma_i}{n_{\text{A},i}} \sum_{m=1}^{n_{\text{A},i}} g_{i,i,k,m}^2 \frac{\|\mathbf{e}_{i,i,k}\mathbf{h}_{i,i,k}^H\|^2}{\|\mathbf{h}_{i,i,k}\|^2}$, and $\frac{\|\mathbf{e}_{i,i,k}\mathbf{h}_{i,i,k}^H\|^2}{\|\mathbf{h}_{i,i,k}\|^2}$ can be expressed by a Gaussian random variable $\tilde{e}_{i,i,k}$ with zero mean and variance $(1 - \lambda_{i,i,k}^2)$ which does not depend on $\mathbf{h}_{i,j,k}$ [34]. Thus, considering also that $\mathcal{E}\{\|h_{i,i,k,l}\|^2\} = 1$ and $\mathcal{E}\{\|\tilde{e}_{i,i,k}\|^2\} = (1 - \lambda_{i,i,k}^2)$, equation (49) becomes:

$$\begin{aligned} \mathcal{E}\{P_x\} \quad &\approx \quad \gamma_i \lambda_{i,i,k}^2 \mathcal{E}\{ \sum_{l=1}^{n_{\text{RRH},i}} g_{i,i,k,l}^2 \} \\ &+ \quad \gamma_i (1 - \lambda_{i,i,k}^2) \mathcal{E}\{\frac{1}{n_{\text{RRH},i}}\} \mathcal{E}\{ \sum_{l=1}^{n_{\text{RRH},i}} g_{i,i,k,l}^2 \} \ , \end{aligned} \tag{51}$$

Equation (51) still contains two unknown terms: $\mathcal{E}\{\frac{1}{n_{\mathrm{A},i}}\} = \mathcal{E}\{\frac{1}{Mn_{\mathrm{R},i}}\}$ and $\mathcal{E}\{\sum_{l=1}^{n_{\mathrm{A},i}} g_{i,i,k,l}^2\}$. For the first one, considering that the number of RRHs in a cluster with unit volume follows a Poisson distribution, i.e. $n_{\mathrm{R},i}/v \sim \pi(\rho_{\mathrm{R}})$, the inverse of $n_{\mathrm{R},i}/v$ will have an inverse Poisson distribution, i.e. exponential distribution:

$$\frac{1}{n_{\mathrm{R},i}/v} = \frac{v}{n_{\mathrm{R},i}} \sim Exp(\rho_{\mathrm{R}}) \ . \tag{52}$$

Thus, the expectation $\mathcal{E}\{\frac{v}{n_{\mathrm{R},i}}\} = \frac{1}{\rho_{\mathrm{R}}}$, i.e. $\mathcal{E}\{\frac{1}{n_{\mathrm{A},i}}\} = \frac{1}{M\rho_{\mathrm{R}}v} = \frac{1}{MN_{\mathrm{R}}}$. Moreover, the second term in (51), $\mathcal{E}\{\sum_{l=1}^{n_{\mathrm{A},i}} g_{i,i,l}^2\}$, stands for the desired signal power given by (11), i.e. without considering delay. The result is given in that case by (19) and (27) for planar and linear case, respectively. Thus, taking the planar deployment as example, replacing $\mathcal{E}\{\frac{1}{n_{\mathrm{A},i}}\}$ and $\mathcal{E}\{\sum_{l=1}^{n_{\mathrm{A},i}} g_{i,i,k,l}^2\}$ by $\frac{1}{MN_{\mathrm{R}}}$ and $\bar{P}_x$, respectively, then we have:

$$\mathcal{E}\{P_x\} = \gamma_i(\lambda_{i,i,k}^2 + \frac{1-\lambda_{i,i,k}^2}{MN_{\mathrm{R}}})\bar{P}_x \ . \tag{53}$$

Similarly, the interference plus noise power can be written as:

$$P_I + \sigma_{i,k}^2 = P_{I,1} + P_{I,2} = \gamma_1 \frac{\sum_{m=1}^{n_{\mathrm{A},i}} g_{i,i,k,m}^2}{n_{\mathrm{A},i}} \sum_{l=1,l\neq k}^{n_{\mathrm{U},i}} \|\check{h}_{i,i,l}\|^2$$
$$+\gamma_1 \sum_{j=1,j\neq i}^{N_c} n_{UE,j} \|\check{h}_{i,j,k}\|^2 \sum_{m=1}^{n_{\mathrm{R},j}} \frac{g_{i,j,k,m}^2}{n_{\mathrm{R},j}} + \sigma_{i,k}^2 \ , \tag{54}$$

where $P_{I,1}$ is the intra-cluster interference which equals to the first item in (54) and $P_{I,2}$ is the inter cluster interference plus noise power corresponding to the addition of second and third items in (54). Note also that $\check{h}_{i,i,k}$ and $\check{h}_{i,j,k}$ are Gaussian random variables with distribution $\mathcal{CN}(0,1)$. Since intra-cluster interference is normally much larger than the inter-cluster interference, we can safely approximate the latter with its expected value, therefore:

$$\mathcal{E}\{\frac{1}{P_I + \sigma_{i,k}^2}\} \approx \mathcal{E}\{\frac{1}{P_{I,1} + \mathcal{E}\{P_{I,2}\}}\} \ , \tag{55}$$

where

$$\begin{aligned}
\mathcal{E}\{P_{I,2}\} &\approx \mathcal{E}\{\gamma_1 \sum_{j=1,j\neq i}^{N_c} n_{UE,j} \|\check{h}_{i,j,k}\|^2 \sum_{m=1}^{n_{\mathrm{RRH},j}} \frac{g_{i,j,k,m}^2}{n_{\mathrm{RRH},j}}\} + \sigma^2 \\
&= \gamma_1 \frac{N_{\mathrm{UE}}}{N_{\mathrm{RRH}}} \mathcal{E}\{ \sum_{j=1,j\neq i}^{N_c} \sum_{m=1}^{n_{\mathrm{RRH},j}} g_{i,j,k,m}^2\} + \sigma_{i,k}^2 \\
&= \gamma_1 \frac{N_{\mathrm{UE}}}{N_{\mathrm{RRH}}} \bar{P}_I + \sigma_{i,k}^2 \ ,
\end{aligned} \tag{56}$$

where $\mathcal{E}\{\sum_{j=1,j\neq i}^{N_c} \sum_{m=1}^{n_{\mathrm{R},j}} g_{i,j,k,m}^2\} = \bar{P}_I$ stands for the interference power derived in (23) and (28) for planar and linear deployment case, respectively.

Regarding intra-cluster interference, replacing $\mathcal{E}\{\sum_{m=1}^{n_{\mathrm{A},i}} g_{i,i,k,m}^2\}$ by $\bar{P}_x$ and considering $\xi_1 = \frac{MN_{\mathrm{R}}}{\gamma_1 \bar{P}_x}$, $P_{I,1}$ will follow a gamma (γ) distribution:

$$P_{I,1} \sim \gamma(N_{\mathrm{U}} - 1, \xi_1) \ , \tag{57}$$

where $N_{\mathrm{U}}-1$ and ξ_1 stand for the shape and rate of the gamma distribution, respectively.

Therefore, (55) is an inverse of a gamma distribution plus a constant $\mathcal{E}\{P_{I,2}\}$ as defined in equation (56). Substituting the pdf of gamma distribution into (55), we have:

$$P_{I,exp} \triangleq \mathcal{E}\{\frac{1}{P_I + \sigma_{i,k}^2}\} = \int_0^\infty u^{N_{\mathrm{U}}-2} e^{-\xi_1 u} \frac{1}{u + P_{I,2}} du \tag{58}$$

By integrating equation (58) in terms of u, we can derive (31). Finally, by substituting (53) and (31) into (48), we can obtain (30).

Appendix C
Proof of (32)

Substituting the second equation of (29) into (8) and considering that the random vector $\mathbf{e}_{i,j,k}$ is independent of $\mathbf{h}_{i,j,k}$, and $\hat{\mathbf{h}}_{i,i,k}\mathbf{w}_{i,k} = 1$, the power of the desired signal can be expressed as:

$$\begin{aligned}
P_x^{ZF} &= \gamma_i \lambda_{i,i,k}^2 \|\mathbf{h}_{i,i,k}\mathbf{G}_{i,i,k}\mathbf{w}_{i,k}\|^2 + \lambda_{i,i,k}^2 \|\mathbf{e}_{i,i,k}\mathbf{G}_{i,i,k}\mathbf{w}_{i,k}\|^2 \\
&= \frac{\gamma_i \lambda_{i,i,k}^2}{[(\mathbf{H}_i\mathbf{G}_{i,i,k}\mathbf{G}_{i,i,k}^H\mathbf{H}_i^H)^{-1}]_k} + \lambda_{i,i,k}^2 \|\mathbf{e}_{i,i,k}\mathbf{G}_{i,i,k}\mathbf{w}_{i,k}\|^2 \ ,
\end{aligned} \tag{59}$$

where notation $[\cdot]_k$ refers to the k-th diagonal element of the matrix and superscript $\{\cdot\}^{ZF}$ is added to differentiate from the MRT precoding solution. By using the approximation of (50), the first item of (59) can be derived as:

$$\frac{1}{[(\mathbf{H}_i\mathbf{G}_{i,i,k}\mathbf{G}_{i,i,k}^H\mathbf{H}_i^H)^{-1}]_k} \approx \frac{\frac{1}{n_{\mathrm{A},i}}\sum_{m=1}^{n_{\mathrm{A},i}} g_{i,i,k,m}^2}{[(\mathbf{H}_i\mathbf{H}_i^H)^{-1}]_k} \ . \tag{60}$$

Similarly, the second item of (59) can be simplified as $\frac{\gamma_i}{n_{\mathrm{A},i}} \sum_{m=1}^{n_{\mathrm{A},i}} g_{i,i,k,m}^2 \sum_{l=1}^{n_{\mathrm{A},i}} \|\tilde{e}_{i,i,k}\|^2$. Thus we can obtain:

$$P_x^{ZF} \approx \gamma_i \frac{\sum_{m=1}^{n_{\mathrm{A},i}} g_{i,i,k,m}^2}{n_{\mathrm{A},i}} (\frac{1}{[(\mathbf{H}_i\mathbf{H}_i^H)^{-1}]_k} + \|\tilde{e}_{i,i,k}\|^2) \ . \tag{61}$$

Since $\mathbf{H}_i$ comprises unit variance and zero mean Gaussian variables, $\mathcal{E}\{1/[(\mathbf{H}_i^H\mathbf{H}_i)^{-1}]_k\} = n_{\mathrm{A},i} - n_{\mathrm{U},i} + 1$ [35], [36]. Thus, considering also $\mathcal{E}\{\sum_{m=1}^{n_{\mathrm{A},i}} g_{i,i,k,m}^2\} = \bar{P}_x$, the expectation of the desired signal power can be written as:

$$\mathcal{E}\{P_x^{ZF}\} = \frac{\gamma_i \lambda_{i,i,k}^2 (MN_{\mathrm{R}} - N_{\mathrm{U}} + 1)}{MN_{\mathrm{R}}} \bar{P}_x + \frac{1-\lambda_{i,i,k}^2}{MN_{\mathrm{R}}} \bar{P}_x \ . \tag{62}$$

Regarding the interference plus noise power, considering that $\hat{\mathbf{h}}_{i,i,k}\mathbf{w}_{i,j} = 0$ for $j \neq k$, we obtain:

$$\begin{aligned}
P_I^{ZF} + \sigma_{i,k}^2 &= \gamma_i \sum_{l=1,i\neq k}^{l=n_{\mathrm{UE},i}} \|(\mathbf{e}_{i,j}\mathbf{G}_{i,j})\mathbf{w}_{i,l}x_{i,l}\|^2 \\
&+\gamma_j \sum_{j=1,j\neq i}^{j=N_c} \sum_{k=1}^{k=n_{\mathrm{UE},i}} \|\bar{\mathbf{h}}_{i,j,k}\mathbf{w}_{j,k}x_{j,k}\|^2 + \sigma^2 \\
&= \gamma_1(1-\lambda_{i,i,k}^2) \frac{\sum_{l=1}^{n_{\mathrm{RRH},i}} g_{i,i,k,l}^2}{n_{\mathrm{RRH},i}} \sum_{k=1,k\neq i}^{n_{\mathrm{UE},i}} \|\check{h}_{i,i,k}\|^2 + P_{I,2} \ .
\end{aligned} \tag{63}$$

Similarly to the MRT precoder solution, using $\mathcal{E}\{P_{I,2}\}$ to replace $P_{I,2}$, the first item of equation (63) has a gamma distribution of $\gamma(N_{\mathrm{U}} - 1, \xi_2)$, where $\xi_2 = \frac{MN_{\mathrm{R}}}{\gamma_1(1-\lambda_{i,i,k}^2)P_x}$. By replacing ξ_1 by ξ_2 in equation (31), we obtain the expectation of interference for ZF-based algorithm. Then, substituting (63) (replace ξ_1 by ξ_2) and (62) into (48), we can derive the output SINR for ZF precoder as given in (32).

REFERENCES

[1] U. Doetsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE," *Bell Labs Technical Journal*, vol. 18, pp. 105–128, May 2013.

[2] D. Wubben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5G signal processing," *IEEE Signal Processing Magazine, Special Issue on 5G Signal Processing*, vol. 31, pp. 35–44, Nov. 2014.

[3] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, "Cloud technologies for flexible 5G radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, May 2014.

[4] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Communications Magazine*, vol. 49, no. 2, pp.102–111, February 2011

[5] L. Li, J. Liu, K. Xiong, P. Butovitsch, "Field test of uplink CoMP joint processing with C-RAN testbed," in *IEEE Communications and Networking in China (CHINACOM)*, pp.753-757, 8–10 Aug. 2012

[6] J. Li, D. Chen, Y. Wang, J. Wu, "Performance evaluation of cloud-ran system with carrier frequency offset," in *IEEE Globecom Workshops (GC Wkshps)*, pp.222–226, 3-7 Dec. 2012

[7] 3GPP TR 36.872, "Small cell enhancements for E-UTRA and E-UTRAN - physical layer aspects," 3GPP, Tech. Rep., Sep. 2013.

[8] 3GPP TR 36.842, "Small cell enhancements for E-UTRA and E-UTRAN - higher layer aspects," 3GPP, Tech. Rep., May 2013.

[9] J.-J. van de Beek, O. Edfors, M. Sandell, S. Wilson, and P. Ola Borjesson, "On channel estimation in OFDM systems," in *IEEE Vehicular Technology Conference*, vol. 2, pp. 815–819, Jul 1995.

[10] S. Colieri, M. Ergen, A. Puri, and A. Bahai, "A study of channel estimation in ofdm systems," in *IEEE Vehicular Technology Conference Fall*, pp. 894–898, 2002.

[11] M. Joham, W. Utschick, and J. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2700–2712, Aug 2005.

[12] M. Matthaiou, C. Zhong, M. McKay, and T. Ratnarajah, "Sum rate analysis of ZF receivers in distributed MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 180–191, February 2013.

[13] D. Kaltakis, M. Imran, and C. Tzaras, "Information theoretic capacity of cellular multiple access channel with shadow fading," *IEEE Transactions on Communications*, vol. 58, no. 5, pp. 1468–1476, May 2010.

[14] S.-R. Lee, S.-H. Moon, J.-S. Kim, and I. Lee, "Capacity analysis of distributed antenna systems in a composite fading channel," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1076–1086, March 2012.

[15] W. Feng, X. Zhang, S. Zhou, J. Wang, and M. Xia, "Downlink power allocation for distributed antenna systems with random antenna layout," in *IEEE Vehicular Technology Conference*, pp. 20-23, Sept 2009.

[16] F. Heliot, R. Hoshyar, and R. Tafazolli, "An accurate closed-form approximation of the distributed MIMO outage probability," *IEEE Transactions on Wireless Communications*, vol. 10, no. 1, pp. 5–11, January 2011.

[17] J. Andrews, F. Baccelli, and R. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, November 2011.

[18] M. Haenggi and R. K. Ganti, *Interference in Large Wireless Networks*. NOW, 2009.

[19] R. Tanbourgi, S. Singh, J.G. Andrews, and F.K. Jondral, "A Tractable Model for Noncoherent Joint-Transmission Base Station Cooperation," *IEEE Transactions on Wireless Communications*, vol.13, no.9, pp. 4959-4973, Sept. 2014

[20] P. Marsch and G. Fettweis, "Rate region of the multi-cell multiple access channel under backhaul and latency constraints," in *IEEE Wireless Communications and Networking Conference*, pp. 830–834, March 2008.

[21] ——, "A framework for optimizing the uplink performance of distributed antenna systems under a constrained backhaul," in *IEEE International Conference on Communications*, pp. 975–979, June 2007.

[22] J. Oueis, E. C. Strinati, and S. Barbarossa, "On the impact of backhaul network on distributed cloud computing," in *IEEE wireless Communications and Networking Conference*, pp. 12–17, April 2014.

[23] X. Ge, K. Huang, C. Wang, X. Hong and X. Yang, "Capacity Analysis of a Multi-Cell Multi-Antenna Cooperative Cellular Network with Co-Channel Interference," *IEEE Transactions on Wireless Communications*, vol.10, no.10, pp.3298-3309, October 2011.

[24] iJOIN Project, "D2.1 - State - of - the - art of and promising candidates for PHY layer approaches on access and backhaul network," Tech. Rep., November 2013. Available: www.ict-ijoin.eu/wp-content/uploads/2014/01/D2.1.pdf.

[25] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*, 2nd ed. Wiley, 2005.

[26] D. S. Sesia, M. M. Baker, and M. I. Toufik, *LTE: The UMTS Long Term Evolution: from Theory to Practice*, 2nd ed. Wiley-Blackwell, 2011.

[27] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.

[28] V. S. Ryaben'kii and S. V. Tsynkov, *A Theoretical Introduction to Numerical Analysis, Section 5.4*. Florida, USA: CRC Press, 2006.

[29] iJOIN Project, "D5.2- final definition of requirements and scenarios," Tech. Rep., November 2014. Available: http://www.ict-ijoin.eu/wp-content/uploads/2012/10/D5.2.pdf.

[30] U. Siddique; H. Tabassum; E. Hossain; I. Dong, "Wireless backhauling of 5G small cells: challenges and solution approaches," *IEEE Wireless Communications*, vol.22, no.5, pp.22–31, October 2015.

[31] X. Ge, H. Cheng, M. Guizani, "5G wireless backhaul networks: challenges and research advances," *IEEE Network*, vol.28, no.6, pp.6-11, Nov. 2014.

[32] L. Yu, W. Liu, and R. Langley, "SINR analysis of the subtraction-based SMI beamformer," *IEEE Transactions on Signal Processing*, vol. 58, pp. 5926–5932, 2010.

[33] L. Zhang, W. Liu, and L. Yu, "Performance analysis for finite sample MVDR beamformer with forward backward processing," *IEEE Transactions on Signal Processing*, vol. 59, pp. 2427 – 2431, May 2011.

[34] H. Q. Ngo, E. Larsson, and T. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, April 2013.

[35] Y. Jiang, M. Varanasi, and J. Li, "Performance analysis of ZF and MMSE equalizers for MIMO systems: An in-depth study of the high SNR regime," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2008–2026, April 2011.

[36] C. Wang, E. Au, R. Murch, W.-H. Mow, R. Cheng, and V. Lau, "On the performance of the MIMO zero-forcing receiver in the presence of channel estimation error," *IEEE Transactions on Wireless Communications*, vol. 6, no. 3, pp. 805–810, March 2007.

[37] I. Agricola, T. Friedrich, "Elementary Geometry," American Mathematical Society, p. 25, 2008.